Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform

Abdelmajid Ben Hamadou, Odile Piton, Héla Fehri

Multilingual Extraction of functional relations between Arabic Named Entities using NooJ platform

Abdelmajid Ben Hamadou (1) Odile Piton (2) and Héla Fehri (3)

(1) MIRACL-University of Sfax, Tunisia.
abdelmajid.benhamadou@isimsf.rnu.tn
(2) SAMM-University of Paris1 Pantheon-Sorbonne, Paris, France.

(2) MIRACL-University of Franche-Comte and University of Sfax, Tunisia
hela.fehri @ fss.rnu.tn

Abstract

The extraction of relation between Named Entities (NE) has become the last few years an interesting research domain. It is very useful for many applications such as Web mining, Information extraction and retrieval, Business intelligence, Automatic databases filing with Entities & types, Questions answering task and document Summarization.
Several works has been performed for relation discovery in texts written in Latin languages and as far as we know, very few works has been done for Arabic language. In this paper, we focus on functional relations between ENAMEX and ORG Arabic Named Entities. The extraction approach is rule based and the implementation is performed using NooJ Platform.

Keywords: Relations between Named Entities, Extraction Process, NooJ transducers, Relation translation

1. Introduction

The extraction of semantic relations between Named Entities (NE) has become the last few years an interesting research domain. It is very useful for many applications such as Web mining, Information extraction and retrieval, Automatic databases filling with Entities & types, Questions answering task and document Summarization.

Relations between Named Entities can be binary (involving two entities) or more complex up to notion of event. There are also several types of relations based on the types of the involved Named Entities.

Several works related to semantic relations discovery has been performed for European languages and especially for English. As far as we know, very few works has been done for Arabic language.

In this paper we propose a rule based relation extraction system for Arabic language. We focus on functional relations between ENAMEX and ORG Named Entities (director, responsible, president…). Functional Relation can be explicit or implicit. Explicit relation is indicated by a special word or sequence of words in the text. Implicit relation is a relation that can be mined from the text using the context.

The extraction process is performed in three steps: Identification of the Named Entities (ENAMED and ORG), Detection of relations between identified Named Entities, Generation of the predicate form representing the relation in Arabic and French. The translation of the relation allows cross lingual information retrieval.

The rest of the paper is organized as follows. We begin by giving an idea about the proposed approaches for extraction of semantic relations between NEs. Then, we address the major challenges posed by extraction of semantic relations between NEs for

Arabic language. After that, we detail our approach and its implementation using NooJ linguistic platform [Silberstein 2005]. Finally, we present the evaluation results obtained from a test corpus.

2. Approaches for NE relation Discovery

Discovering Relations between Named Entities can be done using a linguistic or a numerical or hybrid approach.

The first approach is rule-based, which tried to use syntactic and semantic patterns to capture the corresponding relations by means of manually written linguistic rules. This approach is very interesting for restricted domain and has a good quality of analysis. The major drawback of this approach is the poor adaptability and the poor robustness in handling large-scale or new domain data. This is due to two reasons: rules have to be rewritten for different tasks or when the application is enlarged to different domains and generating rules manually is quite laborious and time consuming (Santos and al., 2010).

The statistical approach (Jun and al., 2009), (Culotta and al., 2006) is based on a learning process from an annotated corpus which can be supervised when the corpus is large, or weakly supervised when the corpus is reduced, or unsupervised if the corpus is not needed. The supervised approach (Miller S. and al., 2000), (Zelenko and al., 2003), (Culotta and al., 2004) and (Kambhatla and al., 2004) costs time and efforts to annotate the corpus, and its performance depends on the size of the corpus. So, to decrease the corpus annotation requirement, some researchers turned to weakly supervised learning approaches, which rely on a small set of initial seed (bootstrap) instead of a large annotated corpus (Agichtein and al., 2000), (Stevenson, 2004). However, the major problem lies in selecting the initial corpus and deciding its "optimal" size.
The unsupervised approach works effectively on high-frequent entity pairs. However, this approach has a limited quality in analysis (Hasegawa and al., 2004).

3. Arabic Named Entities Relations discovery challenges

Arabic Named Entities Relations discovery is not an easy task. In fact, besides the problems related to the Arabic Named Entities recognition (Ben Hamadou and al., 2010), (Shaalan and al. 2009), the extraction of relations between Named Entities poses some specific challenges:

- Discontinuity of the multiple relations concerning the same NE (person).

أ.د. عماد أبو الرب الأمين العام لجمعية كليات الحاسبات - عميد كلية العلوم *Prof. Dr. Imed Abou-Roub the secretary general of organization of computational faculties – Dean of Science Faculty*

The Person Name: *Prof. Dr. Imad Abou-Roub* is concerned by the relation *Secretary-General* and the relation *Dean-of* in the same time. Between the Person Name and the second relation (*Dean-of*) there is a gap.

- Implicit Relations: they are relations that are not directly specified in the text. They are mined from the text using contextual elements.

الدكتور عبد الهادي موسى، اللجنة الشعبية لكلية الطب *Doctor Abd el Hadi Moussa, pupil's committee of medicine Faculty*

The relation between the two NEs: *Doctor Abd el Hadi Moussa* and *pupil's committee of medicine Faculty* is not explicitly indicated by a specific word in the text. In this example the affected relation is *Belong-to*.

- Necessity to use the previous context of the relation in order to know the missing element involved in the relation.
  الأستاذ الدكتور صالح هاشم الأمين العام *Professor Doctor Salah Hachem the secretary general*
  In this example the ORG named entity is absent, but it can be recovered from the right context.

- Interference between implicit relation and discontinuity. In the text we have in the same time implicit relation and discontinuity as shown in this example.

  أ.د. القاسم علي القاسم - كلية الزراعة – جامعة الخرطوم / مساعد الأمين لجمعية كليات الزراعة
  Prof. Dr. El-Kacem Ali El-Kacem – Agricultural Faculty – university of Khartoum / assistant of secretary of faculties' agriculture organization.

  The first implicit relation is *Belong-to* and there is a discontinuity between the Person Name *Prof. Dr El-Kacem Ali El-Kacem* and the Relation *Assistant-of.*

4. The proposed approach

As indicated above, the proposed approach for relation discovery is rule based. It is founded on a balance between grammar and lexical resources. The grammar indicates the composition rules of the lexical components, in order to form the different patterns of functional relations. Patterns are transformed into transducers implemented using NooJ Platform.

The lexical resources correspond to one dictionary for each entity type used by the transducers.

a. Patterns of relations

The approach of extraction functional relations between ORG and ENAMEX Named Entities is based on a notion of linguistic pattern that we transform into rules and transducers.

Patterns are considered as regular expressions integrating different elements representing relations and concerned Named Entities.

These patterns are identified semi-automatically from our journalistic learning corpus using NooJ facilities. Indeed, NooJ allows to identify regular expressions in a corpus such as all First Names with the right and the left contexts, all passengers containing a specific relation with the right and the left context.

In the following, we give a list of the main patterns identified in the learning corpus:

<Pattern 1>:= {<Title>} <PersName> {<P>} <REL> < ORG >
المهندس علي التويجري مدير المجمع الكميائي
Engineer Ali Al-Touijri Director of Chemical group
<Title> ***Engineer*** **المهندس**
<PersName> ***Ali Al-Touijri*** **علي التويجري**
<REL> ***Director-of*** **مدير**
<ORG> ***Chemical group*** **المجمع الكميائي**

<Pattern 2>:= <REL> <ORG > {<P>} {<Title>} <PersName>
مدير المجمع الكميائي: المهندس علي التويجري
Engineer Ali Al-Touijri Director of Chemical group
<Title> ***Engineer*** **المهندس**
<PersName> ***Ali Al-Touijri*** **علي التويجري**
<REL> ***Director-of*** **مدير**
<ORG> ***Chemical group*** **المجمع الكميائي**
<P>:

<Pattern 3>:= {<Title>} <PersName> {<P>} <ORG >
المهندس علي التويجري/ المجمع الكميائي
Engineer Ali Al-Touijri Director of Chemical group
<Title> ***Engineer*** **المهندس**
<PersName> ***Ali Al-Touijri*** **علي التويجري**
<Implicit-REL> ***Belong-To***
<ORG> ***Chemical group*** **المجمع الكميائي**
<P> **/**

<Patern 4>:= <REL> < ***demonym*** -ADJ> {<Title>} <PersName>
الرئيس الأمريكي : باراك أوباما ***US President : Barack Obama***
<REL> ***President*** **الرئيس**
< ***demonym*** -ADJ> ***American*** **الأمريكي**
<PersName> ***Barack Obama*** **باراك أوباما**
<P> **:**

<Patern 5>:= <REL> <Toponym > {<Title>} <PersName>
رئيس الولايات المتحدة : باراك أوباما ***US President : Barack Obama***
<REL> ***President*** **رئيس**
<Toponym > ***US*** **الولايات المتحدة**
<PersName> ***Barack Obama*** **باراك أوباما**
<P> **:**

{ V } Means that the category "V" is optional

<P> Means any punctuation: , /, (-…

b. NooJ grammar

The figure 1 shows the main graph allowing the recognizing of the functional relation between ENAMEX and ORG Named Entities. Each path of this graph represents a different pattern.

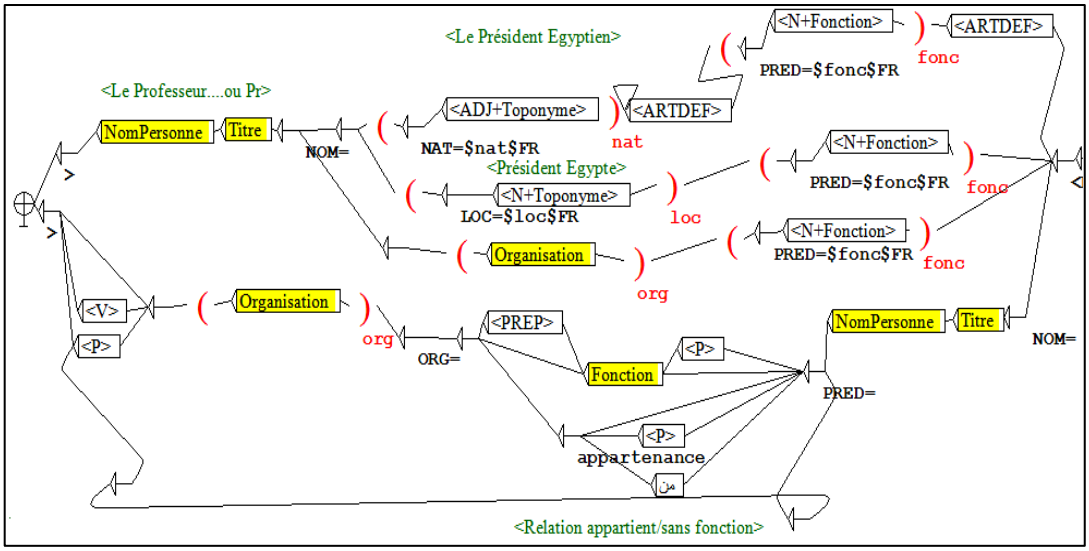

Figure 1: The main graph

The sub graphs:

- *NomPersonne* allows to recognize the Person Names
- *Titre* allows to recognize triggers words preceding the Person Name
- *Organisation* allows the recognition of the ORG Named Entity.
- *Fonction* allows the recognition of the relation between Person Name and ORG Named Entities.

The figure 2: details the *Organisation* graph. It allows the recognition of the ORG Named Entity witch has a complex structure and it includes ENAMEX (Person Name) and LOC (i.e. SN Lieu) Named entities.

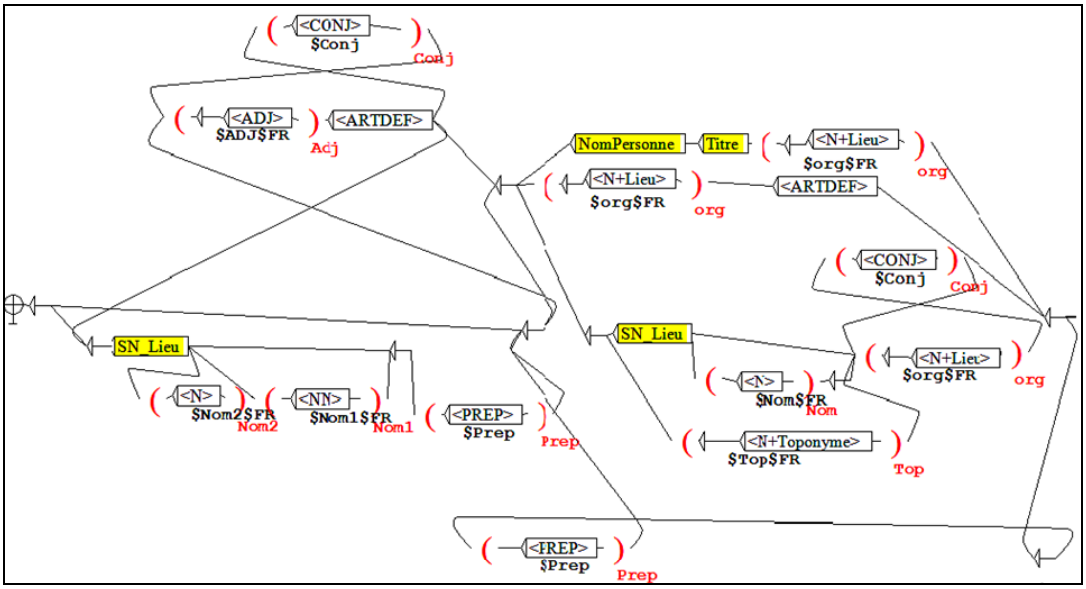

Figure 2: Organisation graph

Beside those graphs, we use a graph for splitting the agglutinated particles (prepositions, definite articles, conjunctions,..) in the analysed Arabic words as shown in this example :
Analysed word: وبالإدارة *et dans l'administration*
Result: Conjonction = و + Préposition = ب + Définite Article = ال + Noun = إدارة

c. NooJ Dictionaries

As mentioned above, the recognition of the relations is based on the following main modules:

- Recognition of the Person Names, including the possible Title,
- Recognition of the ORG Named Entity
- Recognition of the Relation between the indicated Named Entities.

Each module uses its own dictionaries whose entries contain lemmas with the corresponding possible flexional model to generate all derived forms and the corresponding French lemmas.

The recognition of the Person Names module uses the following dictionaries:

- For the Person Names Recognition Step, we build the following dictionaries:
  - Titles Dictionary
  - First Names Dictionary
  - Last Names Dictionary
- For the ORG Recognition module, we build the following dictionaries:
  - Geographical names Dictionary
  - Type institution Dictionary
  - Adjectives Dictionary
- For the Recognition of relations module, we build the following dictionaries:
  - Demonym adjectives Dictionary
  - Functions Dictionary

Table 1 gives extracts of those dictionaries

Extracts of Arabic Dictionaries
 # Title/Functions names
ريدم,N+Fonction+FLX=A1+FR=directeur
كلم,N+Fonction+FLX=كلم+FR=roi
أمين,N+Fonction+FLX=رئيس+FR="Secrétaire"
نائب,N+Fonction+FLX=A1+FR="vice"
مهندس,N+Fonction+Titre+FLX=A1+FR="Ingénieur"
 # names of geographical categories
ةيروهمج,N+FLX=قراق+Cat_Geo+Toponyme+FR=république
ةكلمم,N+FLX=قراق+Cat_Geo+Toponyme+FR=royaume
 # Geographical names
سنوت,N+PR+s+Pays+Toponyme+FR=Tunisie
سنوت,N+PR+s+Ville+Toponyme+FR=Tunis
ضايرا,N+PR+s+Ville+Toponyme+FR=Riyadh
 # Personalities' names
ابزيليإت,N+PR+Perso+f+s+FR=Elisabeth
بيبح ةبيقروب,N+PR+Perso+m+s+FR="Habib Bourguiba"

 # adjectives
يﭭنطو,A+FLX=A1+FR=national
أولمبي,A+FLX=A1+FR=olympique
يﭭدلب,A+FLX=A1+FR=municipal
يﭭلود,A+FLX=A1+FR=international
 # Demonym adjectives
يﭭسنوت,A+Toponyme+FLX=A1+FR=tunisien
يﭭرصم,A+Toponyme+FLX=A1+FR=égyptien
 # Institutions
إتحاد,N+Lieu+FR=union
جمعية,N+Lieu+FR=association
مجمع,N+Lieu+FR=confidération
كلية,N+Lieu+FLX=قارة+FR=faculté
جامعة,N+Lieu+FLX=قارة+FR=université
مكتبة,N+Lieu+FLX=قارة+FR=bibliothèque

Table 1: **Extracts of Arabic dictionaries**

d. Predicate Representation of the Relations

Recognized explicit relations are represented using First Order Logic in the form of a Predicate (i. e. Relation Name) with two arguments: Person Name as the First Argument and the ORG Named Entity as the second Argument. The Name of the relation is the Lemma extracted from the Function Dictionary.

Example:
أمين عام (أبو الرب ، جمعية كليات الحاسبات والمعلومات)
Secretary-General (Abu al-Rub, Association of Colleges of Computing and Information)

For the implicit relation we generate systematically the name "Belong_to" ينتمي_إلى.

Example:
ينتمي_إلى (القاسم علي القاسم ، كلية الزراعة)
Belong_to (Al-Kacem Ali Al-Kacem, Faculty of Agriculture)

e. Translation of the relations

The translation of the recognized relations is done in the perspective of a multilingual assimilation of the analysed texts or documents. And the target language is the French but we can add other languages easily.

So we do not pay a great attention to the quality of the translation process, and we translate only lemmas of the words (i. e., nouns and adjectives) and not their derived forms as they occur in the text. For example:
جامعات Universities is translated as جامعة University,
عامة Générale is translated عام Général (masculine form).
French corresponding lemmas are added to the entries of the different Dictionaries that are used in the recognition process.

This translation, although it is not very good, allows us to question an Arabic text using French language and get answers in Arabic or even in French.

Example:

For the question: *Qui est le Président de l'Université Al-Albeit?* (see figure 4), we get the answer in Arabic: الأستاذ الدكتور الشوقفة and in French: Professeur Docteur Al-Chaouakfa.

5. Experimentation and Evaluation

For the system evaluation, we created a new corpus composed of the learning corpus and Wikipedia pages.

Figure 3 gives a sample execution of the proposed recognition system given by NooJ platform on the indicated corpus. It represents the recognised Relations between Person Names and ORG Named Entities with its corresponding Predicate representation.

Figure 3: A sample of the concordances obtained from the test corpus

Figure 4 gives the recognised relations with their French translation.

الرئيس المصري حسني مبارك/ <REL PRED=président NAT=égyptien NOM= Hosni Moubarak>
الاستاذ الدكتور:نبيل الشو/ <REL NOM= professeur docteur El-Chaouakfa PRED= président ORG= université Aal-Albait >
الدكتور خالد الجيوسي مدير الا/ <REL NOM= docteur El-Jayousi PRED= directeur ORG= administration Général union>
أ.د. عبد القادر بطاح نائب عميد كلية/ <REL NOM= Pr. Dr. Battah PRED= vice doyen ORG= faculté université jordanien>
أ.د. عدلي البلبيسي مدير المجلس العربي/ <REL NOM= Pr. Dr. El-Bilbisi PRED= directeur ORG= conseil arabe>
أ.د.عمار مساعدي ا -/ <REL NOM= Pr. Dr. Msaadi PRED= Secrétaire Général ORG= association scientifique faculté >
فخري جرس أمين عام جمعية كليات التمريض -/ <REL NOM= Jaras PRED= Secrétaire Général ORG= association faculté >
السيد رائد محمد علي مدير جمعية كليات التمريض/ <REL NOM= monsieur Ali PRED= directeur ORG= association faculté >
أ.د. نعمة عازوري الأمين العام للج/ <REL NOM= Pr. Dr. Azouri PRED= Secrétaire Général ORG= association scientifique>
أ.د. سهى ديب أمين عام جمعية -/ <REL NOM= Pr. Dr. Deeb PRED= Secrétaire Général ORG= association faculté sportif>
أ.د. صادق الحايك/ جمعية كليات التربية الر -/ <REL NOM= Pr. Dr. El-Hayek PRED= انتماء ORG= association faculté sportif>
السيد عطية عبد الهادي - عميد كلية السياحة والفنادق –/ <REL NOM= monsieur Abdelhadi PRED= doyen ORG= faculté >
القاسم علي القاسم - كلية الزراعة –/ <REL NOM= El-Kasem PRED= انتماء ORG= faculté >
أ.د. ضياء أحمد القاضي نائ/ <REL NOM= Pr. Dr. El-Khadhi PRED= vice directeur ORG= conseil arabe scientifique>
أبو الرب الأمين العام لجمعية كليات الحاسبات والمعلوما -/ <REL NOM= PRED= Secrétaire Général ORG= association faculté >
أبو الرب الأمين العام لجمعية كليات الحا -/ <REL NOM= Abou-Roub PRED= Secrétaire Général ORG= association faculté >
الصلاح مدير دائرة المجموعات الابداعية و -/ <REL NOM= El-Salah PRED= directeur ORG= circonscription novateur privé>
السهلي أمين عام جمعية كليات الهندسة -/ <REL NOM= El-Shili PRED= Secrétaire Général ORG= association faculté >

Figure 4: A sample of the concordances with their corresponding translation

The evaluation metrics we used for the recognition process, are Recall, Precision and F_measure (2*P*R/(P+R)). Let's remember that the recall measures the quantity of relevant responses of the system compared to the ideal number of responses; Precision is the number of relevant responses of the system among all the responses he gave and the F-measure is a combination of Precision and Recall for penalizing the very large inequalities between these two measures. The values obtained in the evaluation of our work are:

	Precision	***Recall***	***F-measure***
Sports venues	63%	78%	70%

The values obtained can be explained by the fact that our dictionaries are not very large, especially for First Names and Last Names, and also by the fact that the names of organizations are very long and complex.

6. Conclusion

In this paper, we have presented a rule based approach for the extraction of functional relations between ENAMEX and ORG Arabic Named Entities. We particularly highlighted the NE relation Discovery problems. Some of them are specific to the Arabic language. These problems have been largely resolved, but some merit special consideration. The approach was implemented using NooJ platform. We have also given experimentation on a journalistic test corpus. The experimentation and the evaluation results are satisfactory.

As perspectives, we are working on expanding the dictionaries, especially for First Names and Family Names with corresponding translations. Also, we are interested in new types of relations for the economic domain in order to recognize events. Finally we project to integrate our system in a question answering system as a component for factoid questions.

References

Agichtein E. and Gravano L. 2000. Snow-ball: Extracting Relations from Large Plain-text Collections. Proceedings of the Fifth ACM International Conference on Digital Libraries.

Ben Hamadou A., Piton O., Fehri H. 2010. *Recognition and translation Arabic-French of Named Entities: case of the Sport places*, CoRR abs/1002.0481.

César P., Juan P., Isabel S., Paloma M. 2009. The *UC3M team at the Knowledge Base Population task*.

Shaalan Kh. and Raza H. (2009), "*NERA: Named Entity Recognition for Arabic*". Journal Of The American Society For Information Science And Technology, Vol. 60, N° 8, pp: 1652-1663, August 2009

Culotta A., Mccallum A. & Betz J. 2006. *Integrating probabilistic extraction models and data mining to discover relations and patterns in text*. Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Morristown, NJ, USA. +.

Culotta A. and Sorensen J. 2004. *Dependency Tree Kernel for Relation Extraction*. Proceeding of ACL-04.

Jun Z., Zaiqing N., Xiaojiang L., Bo Z. and Ji-Rong W. 2009. *StatSnowball: a Statistical Approach to Extracting Entity Relationships*. 18th international World Wide Web conference (WWW 2009).

Hasegawa T., Sekine S. and Grishman R. 2004. *Discovering Relations among Named Entities from Large Corpora*. Proceeding of ACL-04.

Kambhatla N. 2004. *Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations*. Proceeding of ACL-04, Poster paper.

Miller S., Fox H., Ramshaw L. and Weischedel R. 2000. *A novel use of statistical parsing to extract information from text*. Proceedings of NAACL-00.

Santos D., Mamede N., Baptista J., 2010. *Extraction of Family Relations between Entities*.

Silberstein M. (2005), *"NooJ's dictionaries"*. Actes de la conférence internationale LTC, 2005, Poznan, Pologne.

Silberstein M. (2009), *Syntactic parsing with NooJ.* Proceedings of NooJ 2009, Finite State Language Engineering, Touzeur, Tunisia.

Stevenson M. 2004. *An Unsupervised WordNet-based Algorithm for Relation Extraction*. Proceedings of the 4th LREC workshop "Beyond Named Entity: Semantic Labeling for NLP tasks".

Zelenko D., Aone, C. and Richardella A. 2003. *Kernel Methods for Relation Extraction. Journal of Machine Learning Research*. 2003(2):1083-1106.